Number sequence representation of protein structures based on the second derivative of a folded tetrahedron sequence

Naoto Morikawa (nmorika@genocript.com)

October 7, 2006.

Abstract

A protein is a sequence of amino-acids of length typically less than $1,000$, where there are 20 kinds of amino-acids. In nature, each protein is folded into a well-defined three-dimensional structure, the native structure, and its functional properties are largely determined by the structure. Since many important cellular functions are carried out by proteins, understanding the native structure of proteins is the key to understanding biology at the molecule level.

This paper proposes a new mathematical approach to characterize native protein structures based on the discrete differential geometry of tetrahedron tiles. In the approach, local structure of proteins is classified into finite types according to shape. And one would obtain a number sequence representation of protein structures automatically. As a result, it would become possible to quantify structural preference of amino-acids objectively. And one could use the wide variety of sequence alignment programs to study protein structures since the number sequence has no internal structure.

The programs are available from `http://www.genocript.com`.

Keywords: *protein structure; discrete differential geometry; one-dimensional profile; classification; secondary structure assignment.*

1 Introduction

A protein is a sequence of amino-acids of length typically less than $1,000$, where there are 20 kinds of amino-acids. In nature, each protein is folded into a well-defined three-dimensional structure, the native structure, and its functional properties are largely determined by the structure. Since many important cellular functions are carried out by proteins, understanding the native structure of proteins is the key to understanding biology at the molecule level.

Currently protein structures are characterized by classifications based on structural similarity. But classification is to some extent subjective and there exists a number of classification databases with different organization, such as CATH [3] and SCOP [12]. For example, protein structures are usually described using intermediate structure, such

as α-helix, β-sheet, and turn, which are formed by hydrogen-bonding between distant amino-acids. But there exists no consensus about the assignment of secondary structure, particularly their exact boundaries. (In protein science the intermediate structure is referred to as secondary structure, whereas the amino-acid sequence is primary and the spatial organization of secondary structure elements is tertiary.)

This paper proposes a new mathematical approach to characterize native protein structures based on the discrete differential geometry of tetrahedron tiles [7]. In the approach, local structure of proteins is classified into finite types according to shape. And one would obtain a number sequence representation of protein structures automatically. As a result, it would become possible to quantify structural preference of amino-acids objectively. And one could use the wide variety of sequence alignment programs to study protein structures since the number sequence has no internal structure.

The programs are available from `http://www.genocript.com`.

2 Previous works

An amino-acid sequence has only two rotational freedom per each amino-acid and protein structures was often represented by the two rotational angles, referred as the Ramachandran plot. The torsion angle between Cα atoms were also used to define second structure [6].

[11] proposed a representation of protein structures based on differential geometry. In their method, a protein is represented as broken lines and they defined the curvature and torsion at each point. And [9] described the topology of a protein by 30 numbers inspired by knot theory.

Finally, [10] defined an imaginary cylinder to describe helices, whose axis is approximated by calculating the mean three-dimensional coordinate of a window of four consecutive Cα atoms. And [4] extended the result to define secondary structure based on a number of geometric parameters.

For more information, see [14] which reviews various geometric methods for non-(protein) specialists.

As for one-dimensional profile of protein structures, [1] described every amino acid position in terms of its solvent accessibility, the polarity of its environment and its secondary structure location. And [5] used energy potential to describe amino acid positions.

On the other hand, [13] proposed a "periodic table" constructed from idealized stick-figures as a classification tool, which shifts the classification from a clustering problem to that of finding the best set of ideal structures. And [15] proposed a set of short structural prototypes, "structural alphabet", to encode protein structures.

3 Encoding method

3.1 Differential structure of a tetrahedron sequence

We approximate native protein structures by a particular kind of tetrahedron sequence which satisfies the following conditions (Fig.1(a)): (i) each tetrahedron consists of four

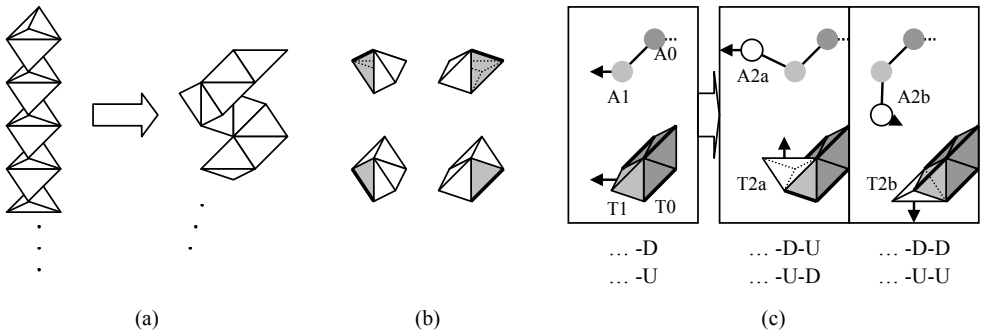


Figure 1: Tetrahedron sequence. (a): Folding. (b): Four directions of a tile (gray). (c): Coding rule (Assignment of the "second derivative").

short edges and two long edges, where the ratio of the length is $\sqrt{3}/2$ and (ii) successive tetrahedrons are connected via a long edge and have the rotational freedom around the edge.

Each tetrahedron of a folded sequence assumes one of the four directions of its short edges, which is determined by the configuration of the previous and next tiles (Fig.1(b)). Then, we can describe change of the direction of tiles, i.e. the "second derivative", by a binary sequence, where either U or D is assigned to each tile. The coding rule is simple: change the value if the direction changes.

For example, suppose that a protein backbone has been approximated up to the previous atom $A0$ by a tetrahedron sequence up to tetrahedron $T0$ (Fig.1(c)). That is, the direction of $T0$ and the position of $T1$ has been determined by the position of the current atom $A1$ (left). Then, there are two candidates $T2a$ and $T2b$ for the position of the next tile $T2$. If the next atom is $A2a$, then $T2a$ is closer to the atom. Thus the next tile assumes the position of $T2a$ and the direction of tetrahedrons changes (middle). In this case assign U to $T1$ if the value of $T0$ is D and assign D to $T1$ otherwise. On the other hand, if the next atom is $A2b$, then the next tile assumes the position of $T2b$ and the direction of tetrahedrons does not change (right). In this case assign D to $T1$ if the value of $T0$ is D and assign U to $T1$ otherwise. At endpoints, choose the tile which is closer to $A1$. (See [7] for the mathematical foundation.)

3.2 5-tile code of proteins

Upon approximation, we consider the position of the centre of amino-acids of proteins. In other words, we identify an amino-acid with the α-carbon atom (Cα) located in its centre. And we allow rotation and translation of tetrahedrons during the folding process to absorb irregularity of actual protein structures. That is, once the direction of tile $T1$ is determined (Fig.1(c)), we rotate $T1$ to make its direction parallel with the direction from $A0$ to $A2$ and translate $T1$ to the position of $A1$.

Fig.2(a), (b), and (c) show approximation of a protein (transferase), whose PDB (Protein Data Bank) ID is 1rkl, by folding only, by folding with rotation, and by folding with rotation and translation respectively.

(a) (b) (c) (d) (e)

Figure 2: Approximation of transferase, PDB ID 1rkl, and others. Broken lines show protein backbones. (a): Folding only. (b): Folding with rotation. (c): Folding with rotation and translation. (d): α-helix (from the 142-th to 154-th amino-acids of protein 1be3). (e): β-sheet (from the 826-th to 838-th amino-acids of protein 1jz7).

To study protein structures, we consider every fragment of an bodd number of amino-acids, say n, contained in a given protein. We start encoding from the middle point amino-acid, say A, and call the obtained binary sequence *n-tile code of amino-acid A*, where A is always assigned value D. (Note that encoding depends on choice of the initial amino-acid.) For example, the middle point amino-acid of the α-helix shown in Fig.1(d) is encoded into 13-tile code $DDDDDUDUDDDDD$ and the middle point amino-acid of the β-sheet shown in Fig.1(e) is encoded into $DDDDDDDDDDDDD$ (all Ds!).

Since it is enough to consider 5-tile codes to detect α-helix, we restrict ourselves on 5-tile codes below. Using 5-tile codes, we obtain 16-valued sequence for each protein. For example, the α-helix of Fig.1(d) is encoded into a 5-tile code sequence of length nine, $HHHHHHHHH$, and the β-sheet of Fig.1(e) is encoded into $SSSTSSSSS$, where H stands for $DUDUD$, S for $DDDDD$, and T for $UUDUU$. See appendix A for the list of 5-tile codes and their examples.

4 Results

4.1 5-tile code assignment of superfolds

It is well known that there exist highly populated families of second structure arrangements, called superfolds [8], shared by a diverse range of amino-acid sequences. And we consider nine superfolds, from which we extract $1,215$ fragments of 5 amino-acids, to study the correspondence between 5-tile codes and secondary structure of proteins, such as helix, sheet, and turn. We use DSSP (Dictionary of Secondary Structure of Proteins) program [2] to assign secondary structure elements to each amino-acid, which calculates energies of hydrogen bonds using a classical electrostatic function.

Table 1(a) shows the frequency distribution of 5-tile codes and their DSSP assignments. This shows that code $DDDDD$ mainly corresponds to sheet, $DUDUD$ mainly to helix, and $UUDUU$ mainly to turn. On the other hand, most of helixes are covered by three codes $DUDUD$, $UUDUD$, and $DUDUU$, most of sheets are covered by $DDDDD$, and most of turns are covered by $UUDUU$, $UUDUD$, and $DUDUU$.

Appendix B shows the spacial distribution of four 5-tile code groups, mainly-helix

Code	*#*	*DSSP assignments (%)*					
	(times)	*Helix*	*Sheet*	*Bridge*	*Turn*	*Bend*	*Else*
DDDDD	534	2	56	2	4	8	27
DUDUD	348	97	0	0	3	0	0
UUDUU	100	0	8	1	34	30	27
UUDUD	81	51	0	0	38	11	0
DUDUU	73	40	0	0	45	11	4
UDDDD	63	3	25	5	6	37	24
Else	16	25	13	0	25	25	13
Total	1215	35	27	1	12	10	16

Table 1: Frequency of 5-tile codes and their DSSP assignments (superfolds).

{DUDUD}, mainly-sheet {DDDDD}, mainly-turn {UUDUU, UUDUD, DUDUU}, and else.

4.2 Frequency distribution of 5-tile codes

Currently protein structures are classified in a hierarchical fashion. For example, the SCOP (structural classification of proteins) database classifies more than $25,000$ proteins into about $3,000$ families manually.

To study the frequency distribution of 5-tile codes, we took one protein for each SCOP family (1.69 release), from which we extracted $1,591,608$ fragments of 5 amino-acids. They corresponds to $612,232$ types of amino-acid sequences, where $26,014$ types are occurred more than nine times.

As table 2(a) shows, 93% of the fragments are covered by top five 5-tile codes. Table 2(b) is concerned with the number of the codes related to one amino-acid sequence, where sequences which occurred more than nine times only are considered. For example, more than 60% of the amino-acid sequences are encoded uniquely and there exist twelve sequences which relate to all of the five codes. Table 2(c) shows their 5-tile code assignments. Note that most of them have a preference for a specific code.

Finally, appendix C gives the preference of amino-acids for 5-tile codes. As you see, amino-acids are grouped into three categories: $DDDDD$-oriented, $DUDUD$-oriented, and others.

5 Discussion

Recently more than a few secondary structure assignment programs are available. But assignments differ from one program to another because of the arbitrariness of the

(a)

Code	*#* *(times)*	*(%)*
DDDDD	641903	(40)
DUDUD	493965	(31)
UUDUU	151282	(10)
UUDUD	101301	(6)
DUDUU	91480	(6)
UDDDD	82585	(5)
DDDDU	12673	(1)
Else	16419	(1)
Total	1591608	(100)

(b)

#Code	*#* *(times)*
5	12
4	209
3	1712
2	7451
1	16052
0	578
Total	26014

(c)

Fragment	*Code assignments* *(times)*
HHHHH	49/ 2/ 8/ 2/ 1/ 4
LEDLG	1/ 1/ 2/37/ 1/ 0
ALAGA	2/ 2/ 3/10/ 3/ 2
DPDLV	2/12/ 3/ 1/ 2/ 1
ALLSD	6/ 3/ 2/ 2/ 2/ 2
GSLGS	2/ 2/ 1/10/ 1/ 0
LPGIG	3/ 1/ 9/ 2/ 1/ 0
HSAGV	1/ 2/ 3/ 1/ 6/ 2
GSLGA	1/ 2/ 2/ 2/ 6/ 1
DEGTG	1/ 1/ 2/ 4/ 1/ 4
GAAGG	4/ 2/ 2/ 3/ 2/ 0
ADAAD	3/ 2/ 2/ 1/ 2/ 0

Table 2: Statistics of 5-tile codes. (a): Frequency of 5-tile codes. (b): Frequency of amino-acid sequences which relate to a given number of the top five 5-tile codes. (Sequences which occurred more than nine times only are considered) (c): 5-tile code assignments of the amino-acid sequences which relate to all of the top five 5-tile codes. Sequences are given by one-letter code of amino-acids and the figures show the frequency of $DDDDD$/$DUDUD$/$UUDUU$/$DUDUU$/$UUDUD$/Else.

definition of secondary structure, particularly its exact boundaries. According to [4], the degree of disagreement between two different assignments could reach almost 20%.

On the other hand, our approach dose not require any pre-defined secondary structure. Instead, it classifies local structure according to shape. For example, in the case of 5-tile coding, local structure is grouped into 16 types of elements (See appendix A). Though it can neither distinguish three types of helixes from one another nor describe global features directly, such as hydrogen bonding, it can detect secondary structure to same extent (See appendix B). That is, it allows a comprehensive description of, not only specific (secondary) structure, but also arbitrary local structure. And it becomes possible to quantify structural preference of amino-acids objectively (See appendix C). It could be used to characterize binding sites of drugs and the active site of enzymes, too.

Moreover, since our method encodes a protein into a number sequence without any structure, it is easy to analyse the (number sequence) profile and we could use the wide variety of sequence alignment programs to compare protein structures. And the incorporation of structural information should lead to more powerful protein structure predictions because structure is much more strongly conserved than sequence during evolution. For example, in the case of 5-tile coding, more than 60% of amino-acid sequences are uniquely encoded. Thus, they can be substituted with 5-tile codes when

one considers plausible structures of a given amino-acid sequence.

6 Conclusions

This paper proposes a new mathematical approach to characterize native protein structures based on the discrete differential geometry of tetrahedron tiles [7]. In the approach, local structure of proteins is classified into finite types according to shape. And we have obtained a comprehensive description of, not only specific (secondary) structure, but also arbitrary local structure elements. As a result, one could quantify structural preference of amino-acids objectively. And one could use the wide variety of sequence alignment programs to study protein structures since the one-dimensional profile of the assignment is given as a simple number sequence.

References

[1] Bowie, J.U., Luthy, R. & Eisenberg, D. 1991 A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164-169.

[2] Kabsch, W. & Sander, C. 1983 Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-2637.

[3] Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., & Thornton, J.M. 1997 CATH- A Hierarchic Classification of Protein Domain Structures. *Structure.* **5**, 1093-1108.

[4] Cubellis, M.V., Caulliez, F., & Lovell, S.C. 2005 Secondary structure assignment that accurately reflects physical and evolutionary characteristics. *BMC Bioinformatics* **6** (Suppl 4), S8.

[5] Jones, D.T., Taylor, W.R. & Thornton, J.M. 1992 A new approach to protein fold recognition. *Nature* **358**, 86-89.

[6] Levitt, M. 1983 Protein folding by restrained energy minimization and molecular dynamics. *J. Mol. Biol.* **170**, 723-64.

[7] Morikawa, N. 2006 Discrete differential geometry of triangle tiles and algebra of closed trajectories. ArXiv: math.CO/0607051.

[8] Orengo, C.A., Jones, D.T. & Thornton, J.M. 1994 Protein superfamilies and domain superfolds. *Nature* **372**, 631-634

[9] Rogen, P & Fain, B. 2003 Automatic classification of protein structure by using Gauss integrals. *Proc. Natl. Acad. Sci.* **100**, 119-124.

[10] Richardson, J.S. & Richardson, D.C. 1988 Amino acid preferences for specific locations at the ends of alpha helices. *Science* **240**, 1648-1652.

[11] Rackovsky, S. & Scheraga, H.A. 1978 Differential Geometry and Polymer Conformation. 1. *Macromolecules* **11**, 1168-1174.

[12] Murzin A. G., Brenner S. E., Hubbard T., & Chothia C. 1995 SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.

[13] Tayler, W.R. 2002 A 'periodic' table for protein structre. *Nature* **416**, 657-660.

[14] Taylor, W.R. & Aszodi, A. 2005 Protein Geometry, Classification, Topology and Symmetry: A computational analysis of structure. IOP Publishing Ltd., London.

[15] Tyagi, M., Sharma, P., Swamy, C.S., Cadet, F., Srinivasan, N., de Brevern, A.G., Offmann, B. 2006 Protein Block Expert (PBE): a web-based protein structure analysis server using a structural alphabet. *Nucleic Acids. Res.* **34**, W119-W123 (Web server issue).

A 5-tile code list

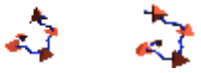
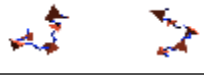
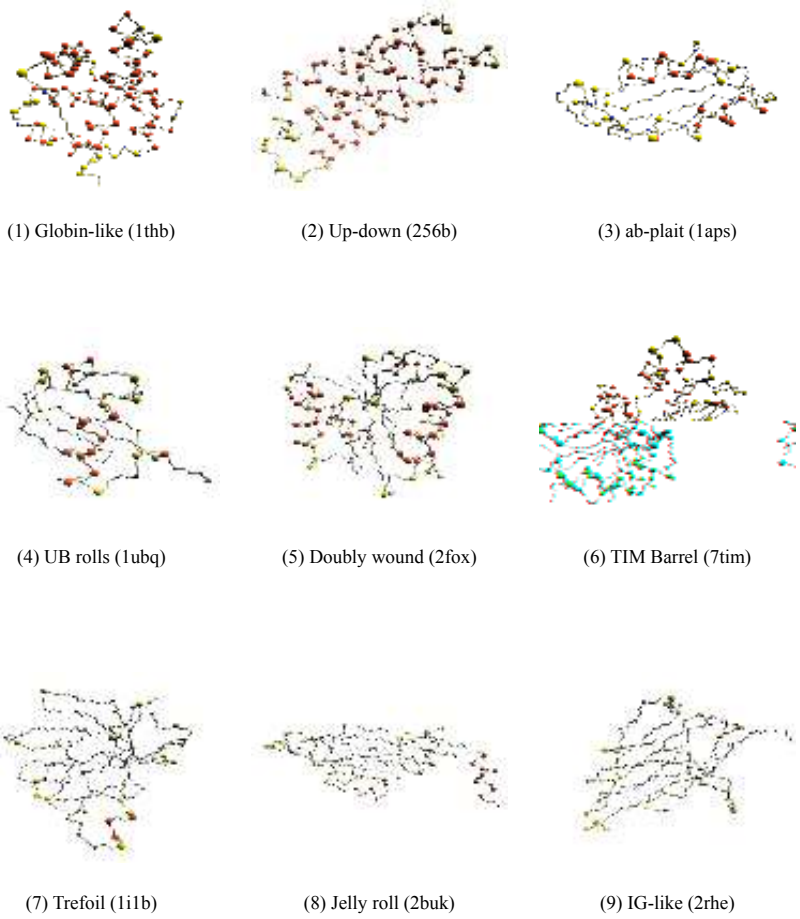
5-tile code	Examples (amino-acid fragments)	5-tile code	Examples (amino-acid fragments)
DDDDD		UUDDD	
DUDUD		DDDUU	
UUDUU		DUDDD	
UUDUD		UDDUU	
DUDUU		DUDDU	
UDDDD		UUDDU	
DDDDU		UDDUD	
UDDDU		DDDUD	

Table 3: 5-tile code list. Broken lines show protein backbones.

B Spacial distributions of 5-tile codes



(1) Globin-like (1thb) (2) Up-down (256b) (3) ab-plait (1aps)

(4) UB rolls (1ubq) (5) Doubly wound (2fox) (6) TIM Barrel (7tim)

(7) Trefoil (1i1b) (8) Jelly roll (2buk) (9) IG-like (2rhe)

Figure 3: 5-tile code distributions of nine superfolds. Broken lines show protein backbones and Red balls are the Cα atoms of the amino-acids of mainly-helix code $DUDUD$, yellow balls mainly-turn codes $\{UUDUU, DUDUU, UUDUD\}$, and small blue balls mainly-sheet code $DDDDD$.

C 5-tile code preference of amino-acids

Amino-acids	*#*	*5-tile codes (%)*					
	(times)	*DDDDD*	*DUDUD*	*UUDUU*	*UUDUD*	*DUDUU*	*Else*
VAL (V)	118030	54	28	4	5	2	7
ILE (I)	92900	49	33	4	5	3	7
SER (S)	90250	44	26	8	7	7	8
THR (T)	89111	49	23	8	5	6	8
PHE (F)	62974	45	30	6	5	6	8
TYR (Y)	54210	47	28	6	5	6	8
HIS (H)	36063	42	27	9	5	7	8
TRP (W)	21091	42	34	5	8	5	6
CYS (C)	20460	51	25	8	3	5	8
LEU (L)	144979	35	41	6	6	6	6
ALA (A)	133378	31	45	6	7	6	5
GLU (E)	109542	29	43	7	8	7	5
LYS (K)	91565	34	35	9	7	7	7
ARG (R)	86253	36	37	8	6	6	6
GLN (Q)	58778	32	40	8	5	8	7
MET (M)	32351	36	41	6	5	6	7
GLX (Z)	46	28	52	13	4	0	0
GLY (G)	120167	43	14	30	4	3	5
ASP (D)	92614	35	26	16	6	10	7
PRO (P)	71132	55	8	10	23	1	2
ASN (N)	65487	35	24	19	3	11	8
ASX (B)	28	43	36	14	0	4	0
Total	1591409	40	31	10	6	6	6

Table 4: Preference of amino-acids for 5-tile codes.